

94-363



ОБЪЕДИНЕННЫЙ
ИНСТИТУТ
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ
ДУБНА

P10-94-363

А.Ю.Бонюшкина, П.В.Зрелов, В.В.Иванов

ОБ ОДНОМ ВЫБОРЕ ВХОДНЫХ ДАННЫХ
ДЛЯ МНОГОСЛОЙНОГО ПЕРЦЕПТРОНА

Направлено в журнал «Computer Physics Communications»

1994

Введение

В последнее время искусственные нейронные сети находят все большее применение в экспериментальной физике высоких и промежуточных энергий при построении систем, предназначенных для быстрого отбора событий в ходе реального эксперимента [1]. Кроме того, наряду с традиционными методами математической статистики, они с успехом используются в задачах массовой обработки накопленной информации [2].

Одна из таких задач состоит в классификации отдельных событий, представляемых эмпирическими выборками конечного объема в соответствии с их принадлежностью к одному из парциальных распределений, образующих анализируемый спектр. В работе [4] рассмотрены два различных подхода для решения указанной проблемы с использованием одного из типов искусственных нейронных сетей – многослойного перцептрона (MLP¹) [3, 1] и непараметрических критериев согласия ω_n^k [5]. Критерии ω_n^k относятся к классу критериев, основанных на эмпирической функции распределения, построенной на упорядоченной выборке исследуемых величин (вариационном ряде). В работе [4] предложено использование вариационного ряда в качестве входных данных также и для MLP.

В настоящей работе анализируются результаты обучения многослойного перцептрона для различных представлений входных данных, исследуются причины, приводящие к эффекту "мгновенного" обучения нейронной сети и повышения ее мощности для данных в форме вариационного ряда, обсуждается проблема сокращения числа нейронов в скрытом слое без увеличения ошибки распознавания.

¹MLP – Multi-Layer Perceptron.

1 Классификация данных при помощи MLP

Общие принципы работы многослойного перцептрона и его применение для задач классификации удобно рассмотреть на примере из работы [1]. Пусть требуется построить классификатор событий, принадлежащих либо классу "а", либо классу "б" (см. рис.1). Этой цели отвечает разделяющая функция следующего вида:

$$D = \Theta[\Theta(a_1x + b_1y + c_1) + \Theta(a_2x + b_2y + c_2) + \Theta(a_3x + b_3y + c_3) - 2], \quad (1)$$

где $\Theta(x - x')$ – пороговая функция, равная 0 при $x < x'$ и 1 при $x \geq x'$.

Параметры a_i, b_i и $c_i, i = 1, \dots, 3$ подобраны так, что функция (1) принимает значение, равное 1 в области, включающей границы класса "б", и значение, равное 0, в остальной области.

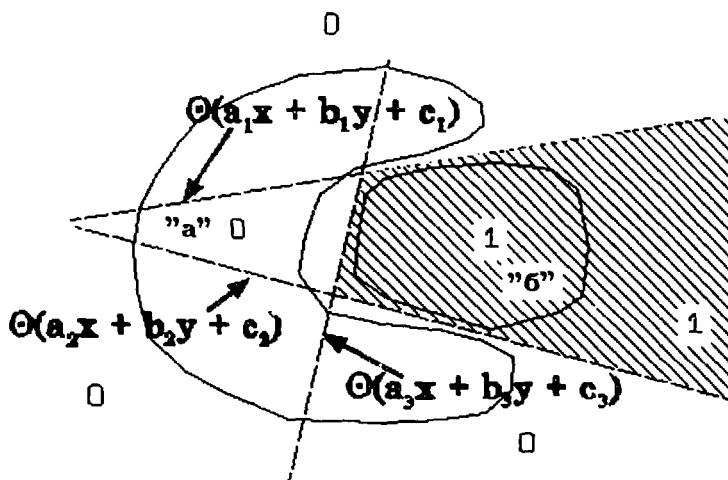


Рис.1: Области событий, отвечающие двум разным классам: "а" или "б"

На рис. 2 изображена возможная схема классификатора, реализующая функцию (1). На входы x и y подаются случайные величины, отвечающие текущему событию. Эти величины умножаются на коэффициенты a_i и $b_i, i = 1, \dots, 3$, определяющие веса линий, связывающих входы схемы с первым слоем пороговых дискриминаторов (ПД). На входы ПД, кроме того, подаются пороги c_1, c_2 и c_3 , отвечающие за параллельный сдвиг соответствующей разделительной линии на плоскости. Сигналы с выходов первых ПД умножаются на новые коэффициенты, равные в нашем примере 1, и вместе с порогом, равным -2 , подаются на выходной ПД.

Рассмотренная схема дискриминирующей функции является упрощенной моделью трехслойной нейронной *feed-forward* сети, в которой роль нейронов выполняют пороговые дискриминаторы. Указанная сеть содержит два входных

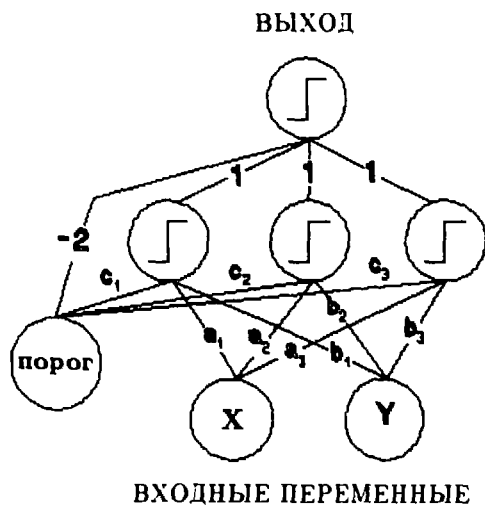


Рис.2: Схема дискриминирующей функции

нейрона, три нейрона в скрытом слое, выходной нейрон и нейрон, задающий пороги. Коэффициенты a_1, b_1, \dots образуют матрицу весов сети. В многослойном перцептроне каждый j -ый нейрон осуществляет преобразование $t_j = g(\sum_i \omega(j, i)t_i + \theta_j)$, где t_j – выходной сигнал рассматриваемого нейрона, $\omega(j, i)$ – вес связи между i -м и j -м нейронами, g – переходная функция (в рассмотренном выше примере она совпадает с функцией Θ), а θ_j – соответствующие пороги. Обычно в нейронной сети вместо ступенчатой функции используют переходную функцию сигмоидного типа, в частности:

$$g(x) = \tanh(x). \quad (2)$$

Такая функция существенно расширяет возможности нейронной сети, позволяя проводить коррекцию весов² межнейронных связей и строить гладкую границу, разделяющую альтернативные распределения.

В общем случае разделения классов в n -мерном пространстве решающая граница приобретает форму сложной гиперповерхности, устанавливаемой MLP в процессе итерационной процедуры, называемой *обучением* [7]. Очевидно, что преобразование исходных классов, приводящее к упрощению характера решающей границы, сокращению ее протяженности, должно вести в целом к ускорению процесса обучения и уменьшению ошибки классификации.

²Для этих целей обычно используется алгоритм *back-propagation*, требующий вычисления первых производных от функции $g(x)$ (см., например, [6]).

2 Выбор входных данных для MLP

Рассмотрим задачу классификации событий, которые представлены выборками x_i , $i = 1, \dots, n$ объема от $n = 2$ до 9, принадлежащими гауссовским распределениям с совпадающими средними значениями $N(0, 1)$ и $N(0, 0.3)$ ³[4]. Будем считать, что события, отвечающие распределению $N(0, 1)$ относятся к I типу, а $N(0, 0.3)$ – ко II типу. Для классификации событий использовалась трехслойная *feed-forward* сеть из пакета JETNET 2.0 [7], содержащая n входных нейронов, скрытый слой из 16 нейронов и один выходной нейрон. Все нейроны имели одну и ту же переходную функцию вида (2). При обучении MLP для событий I типа сигнал на выходе сети полагался равным -1 , а для событий II типа $+1$. Классификации было подвергнуто около 8000 событий I типа и 2000 событий II типа. На рис.3 представлено суммарное распределение для смеси выборок обоих распределений.

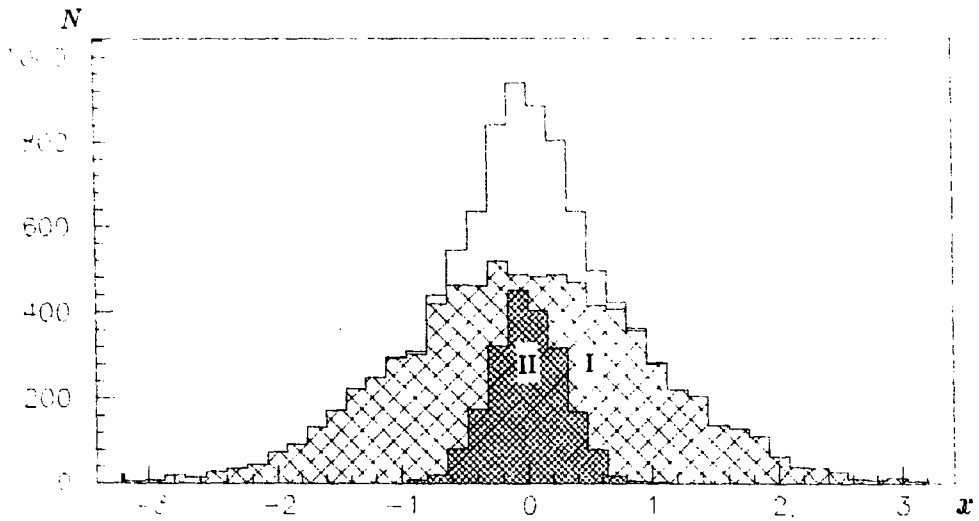


Рис.3: Суммарное распределение для смеси выборок из двух распределений $N(0, 1)$ и $N(0, 0.3)$; вклады от I и II распределений выделены соответствующей штриховкой

Все события были разбиты на две равные по количеству группы. Первая группа использовалась для обучения сети, вторая – для её тестирования. Идентификация событий осуществлялась по величине выходного сигнала y – если она не превышала заданное пороговое значение y_t , то рассматриваемое событие приписывалось I типу, в противном случае – II типу.

В каждом примере выполнялось необходимое для настройки на решаемую задачу количество обучающих циклов (эпох). При этом в каждом цикле на вход сети подавались одни и те же данные, а по его завершении проводилась коррекция весов межнейронных связей. По окончании обучения фиксировались веса и пороги-смещения нейронов.

³ $N(a, \sigma)$ – гауссовское распределение со средним a и дисперсией σ^2

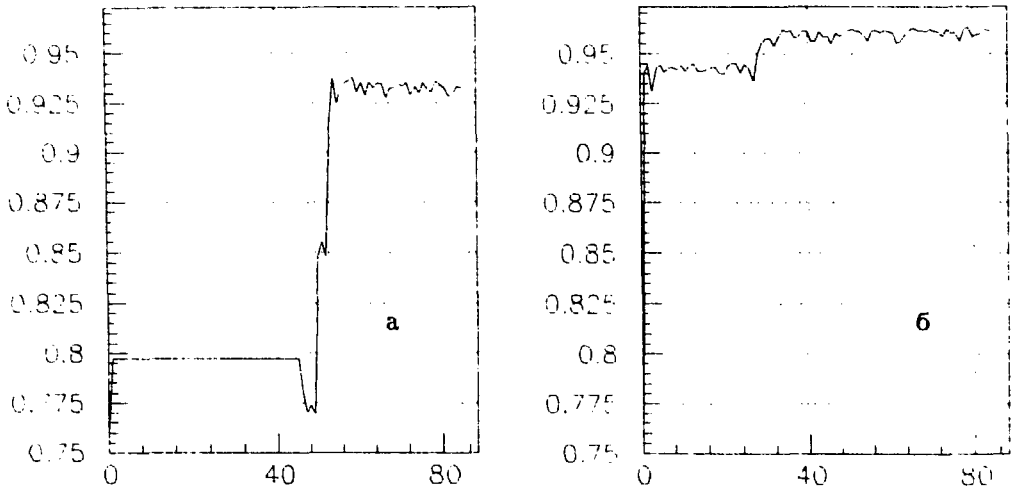


Рис.4: Кривая вероятности распознавания событий нейронной сетью в зависимости от номера обучающего цикла для $n = 5$, когда входные данные представлены: а) выборками случайных величин, б) выборками, преобразованными к вариационному ряду

Рассмотрим случай, когда на вход нейронной сети подаются выборки случайных величин x_i . Для выборки объема $n = 2$ сеть начинает идентифицировать события, принадлежащие разным распределениям, только после 485 эпохи. С увеличением n для достижения хорошего уровня распознавания требуется все меньшее количество тренировочных циклов. На рис.4 а в качестве примера приведена кривая вероятности распознавания событий в зависимости от номера обучающего цикла для $n = 5$.

Перейдем от исходной выборки к упорядоченной выборке \tilde{x}_i (вариационному ряду), образованной из элементов x_i :

$$\tilde{x}_1 < \tilde{x}_2 < \dots < \tilde{x}_n, \quad i = 1, \dots, n. \quad (3)$$

На рис.4б приведена соответствующая кривая вероятности распознавания событий для случая, когда на вход сети подаются упорядоченные выборки ($n = 5$). Характер кривой говорит о том, что обучение сети происходит практически мгновенно⁴. Сопоставление зависимостей рис.4а и б показывает, что после завершения процесса обучения вероятность распознавания событий для варианта с упорядоченной выборкой оказывается выше.

На рис.5а представлены спектры величин y выходных сигналов сети, полученные в результате обработки классифицируемых событий: "чистая" гистограмма

⁴Отметим, что для выборок \tilde{x}_i объема $n=2$ этот момент наступает уже на 5-ой эпохе в отличие от 485-ой эпохи для случайной выборки x_i .

отвечает выборкам \tilde{x}_i из первого распределения, а штрихованная гистограмма из второго распределения. На рис.56 для событий II типа приведена зависимость функции накопленной вероятности $F(y_t) = \text{Pr}\{y < y_t\}$, равной доле неверно классифицируемых событий этого типа (ошибка второго рода β), и зависимость $1 - F(y_t) = \text{Pr}\{y > y_t\}$ для событий I типа (ошибка первого рода α).

Сопоставление указанных зависимостей для двух вариантов входных данных показывает, что в точке пересечения кривых $\alpha = \beta$ уровень распознавания составляет:

а) для неупорядоченных входных данных -- 0.929,

б) для упорядоченных входных данных -- 0.960,

т.е. во втором случае эффективность классификации событий нейронной сетью оказывается выше.

Таким образом, в случае использования в качестве входных данных для MLP вариационного ряда существенно сокращается время обучения сети, а также возрастает ее мощность.

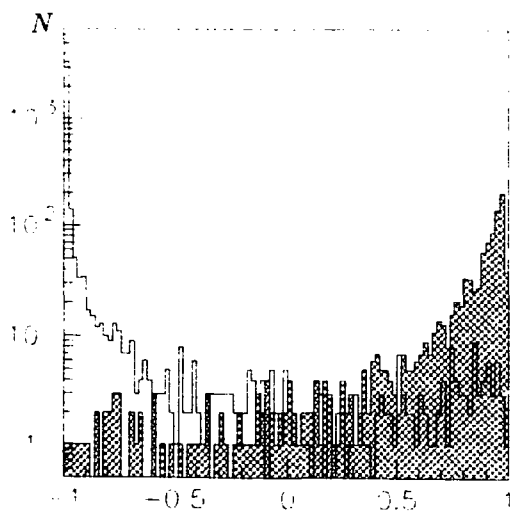


Рис.5а: Спектры величин выходных сигналов сети, полученные в результате обработки классифицируемых событий: "чистая" гистограмма - выборки из I распределения, штрихованная гистограмма - выборки из II распределения: в качестве входных данных используются выборки упорядоченных случайных величин \tilde{x}_i ;

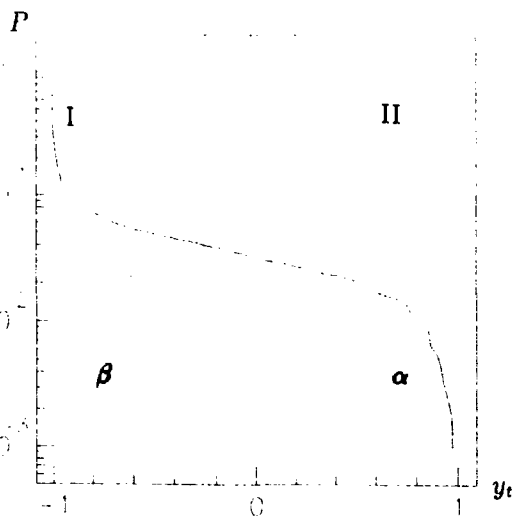


Рис.5б: Кривые зависимости величины накопленной вероятности $F(y_t) = \text{Pr}\{y < y_t\}$ в случае выборки объемом $n = 5$ из II распределения и зависимости $1 - F(y_t)$ для I распределения: в качестве входных данных используются выборки упорядоченных случайных величин \tilde{x}_i ;

3 Обсуждение результатов

При байесовской классификации с минимальным уровнем ошибки для разделения ν классов ω_j , $j = 1, 2, \dots, \nu$ событий, подчиняющихся многомерным гауссовским распределениям с ковариационными матрицами Σ_j и векторами средних значений $\vec{\mu}_j$, используются разделяющие функции следующего вида (см., например, [11]):

$$g_j(\vec{x}) = -\frac{1}{2}(\vec{x} - \vec{\mu}_j)' \Sigma_j^{-1} (\vec{x} - \vec{\mu}_j) - \frac{1}{2} \ln |\Sigma_j| + \ln P(\omega_j), \quad (1)$$

где $P(\omega_j)$ – априорные вероятности, отражающие первоначальное знание о соотношении разделяемых классов, $|\Sigma_j|$ – определитель ковариационной матрицы.

Исследуемой в данной работе задаче отвечает случай двух классов ω_1 и ω_2 . Для простоты положим $P(\omega_1) = P(\omega_2)$, что означает выбор соотношения между классами 1 : 1 и позволяет пренебречь членом $\ln P(\omega_j)$ в выражении (1). Кроме того, для рассматриваемой задачи предполагается, что ковариационные матрицы Σ_j имеют диагональную структуру, т.е. $\Sigma_j = \sigma_j^2 I$, $j = 1, 2$, где I – единичная матрица.

Поверхность решений, удовлетворяющая условию

$$g_1(\vec{x}) = g_2(\vec{x}),$$

определяется следующим уравнением

$$\sum_{i=1}^n x_i^2 + 2 \sum_{i=1}^n x_i b_i + c = 0, \quad (2)$$

где

$$b_i = \frac{\sigma_2^2 \mu_{1i} - \sigma_1^2 \mu_{2i}}{\sigma_1^2 - \sigma_2^2}, \quad c = \frac{\sigma_1^2}{\sigma_1^2 - \sigma_2^2} \vec{\mu}_2^2 - \frac{\sigma_2^2}{\sigma_1^2 - \sigma_2^2} \vec{\mu}_1^2 + \frac{2\sigma_1^2 \sigma_2^2}{\sigma_1^2 - \sigma_2^2} n \ln \frac{\sigma_2}{\sigma_1}.$$

Следуя классификации поверхностей второго порядка в n -мерном евклидовом пространстве [13], можно показать, что уравнение (2) определяет гиперсферу с центром в точке $(-b_1, -b_2, \dots, -b_n)$ и радиусом $R = \sqrt{-|A_1|}$, где

$$A_1 = \begin{pmatrix} 1 & 0 & \dots & 0 & b_1 \\ 0 & 1 & \dots & 0 & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & b_n \\ b_1 & b_2 & \dots & b_n & c \end{pmatrix}. \quad (3)$$

Детерминант матрицы (3) определяется равенством $|A_1| = -\sum_{i=1}^n b_i^2 + c$, и, следовательно, радиус гиперсферы равен

$$R = \frac{\left\{ \sigma_1^2 \sigma_2^2 \|\vec{\mu}_1 - \vec{\mu}_2\|^2 - 2n\sigma_1^2 \sigma_2^2 (\sigma_1^2 - \sigma_2^2) \ln \frac{\sigma_2}{\sigma_1} \right\}^{1/2}}{|\sigma_1^2 - \sigma_2^2|}. \quad (4)$$

Для плоского случая ($n = 2$) с $\vec{\mu}_i = 0$, $i = 1, 2$ решающая поверхность приобретает характер окружности с центром $(0, 0)$ и радиусом

$$R = 2 \sigma_1 \sigma_2 \left\{ \frac{\ln \frac{\sigma_1}{\sigma_2}}{\sigma_1^2 - \sigma_2^2} \right\}^{1/2}. \quad (5)$$

На рис.6а представлены области, содержащие 95% событий, образуемых случайными выборками объема $n = 2$ из распределений $N(0, 1)$ и $N(0, 0.3)$. Отмечена также решающая окружность, радиус которой вычислен по формуле (5) для случая $\sigma_1 = 1$, $\sigma_2 = 0.3$ и составляет $R = 0.6901$. Классификация относительно этой границы обеспечивает предельный уровень распознавания 0.8586.

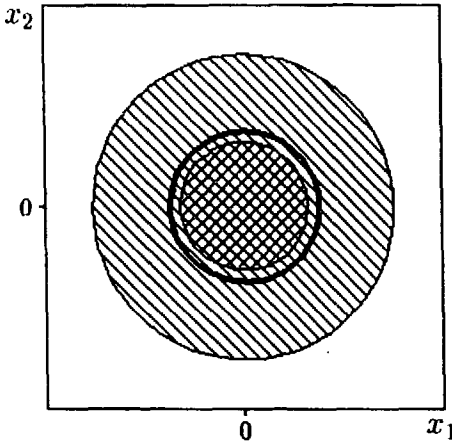


Рис.6а: Доверительные области (95%) для двумерных распределений, образуемых случайными выборками объема $n = 2$ из $N(0, 1)$ и $N(0, 0.3)$. Области, относящиеся к разным классам, выделены штриховкой. Жирной линией отмечена байесовская решающая окружность

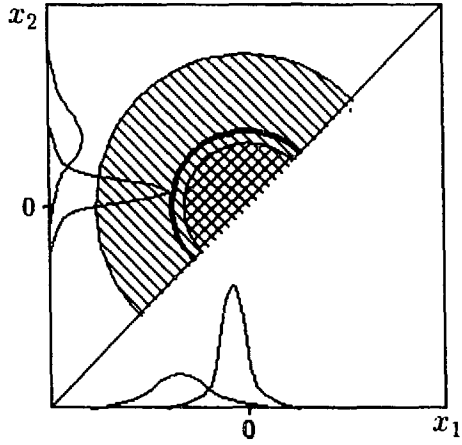


Рис.6б: Доверительные области (95%) для двумерных распределений, образуемых упорядоченными выборками объема $n = 2$ из $N(0, 1)$ и $N(0, 0.3)$. Области, относящиеся к разным классам, выделены штриховкой. Жирной линией отмечена байесовская решающая граница. Схематически представлены также распределения, определяющие входы нейронов

Рассмотрим что происходит с байесовской границей в случае $n = 2$ при переходе от случайной к упорядоченной выборке. Геометрически процедура упорядочивания означает отражение точек плоскости, для которых $x > y$, относительно прямой $x = y$. При этом плотность вероятности результирующего двумерного распределения в области $x > y$ равна нулю, а в области $x < y$ представляет собой удвоенную гауссову плотность. Поэтому в области $x < y$ выражение для разделяющей функции не меняется и, решающая кривая представляет собой по-

дуокружность того же радиуса R (рис.6б). На обоих рисунках области, относящиеся к разным классам событий, выделены штриховкой.

Рассмотренное преобразование приводит к тому, что решающая граница перестает быть замкнутой, и это значительно упрощает поиск минимума соответствующего функционала в процессе обучения сети (см., например, [8]), увеличивая тем самым скорость этого обучения. Другой причиной являются сокращение "площади" границы, приводящей к уменьшению ошибки в процессе настройки сети.

Рассмотрим теперь, как статистически меняется ситуация при переходе от случайной к упорядоченной выборке. В этом случае элементы x_m вариационного ряда $x_1 < x_2 < \dots < x_m < \dots < x_n$ перестают быть независимыми и принадлежать одному и тому же распределению. Их плотности описываются выражением (см., например, [9]):

$$[B(m, n - m + 1)]^{-1} [F(x)]^{m-1} [1 - F(x)]^{n-m} f(x), \quad (6)$$

где $F(x)$ - распределение, которому принадлежит эмпирическая выборка, $f(x)$ - его плотность, а B - биномиальные коэффициенты.

На рис.6б схематически представлены плотности распределений для элементов упорядоченных выборок ($m = 1$ и 2) из распределений $N(0, 1)$ и $N(0, 0.3)$, в соответствии с которыми события разделяемых классов подаются на входы нейронной сети. Хорошо видно, что эти распределения имеют заметный сдвиг по среднему. Величина этого сдвига может быть вычислена с использованием выражений для центральных моментов распределения (6), приведенных в работе [10]:

$$\begin{aligned} \mu'_1 = x_p - \frac{f'}{2f^3} \cdot \frac{pq}{n+2} + \frac{3f'^2 - ff''}{6f^5} \cdot \frac{2pq(q-p)}{(n+2)(n+3)} + \\ + \frac{10ff'f'' - f^2f''' - 15f'^3}{24f^7} \cdot \frac{3p^2q^2N + 3pq(2-5pq)}{(n+2)(n+3)(n+4)} + \dots, \end{aligned} \quad (7)$$

$$\begin{aligned} \mu_2 = \frac{1}{f^2} \cdot \frac{pq}{n+2} + \frac{f'}{f^4} \cdot \frac{2pq(q-p)}{(n+2)(n+3)} + \\ + \left[\frac{5f'^2}{4f^6} - \frac{f''}{3f^5} \right] \cdot \frac{3p^2q^2N + 3pq(2-5pq)}{(n+2)(n+3)(n+4)} - \frac{f'^2}{4f^6} \cdot \frac{p^2q^2}{(n+2)^2} + \dots, \end{aligned} \quad (8)$$

где $x_p = F^{-1}(p)$, $f^{(i)} = f^{(i)}(x_p)$, $p = m/(n+1)$, $q = 1 - p$. Формулы (7) и (8) использовались для вычислений средних значений и дисперсий соответствующих распределений для разных объемов эмпирической выборки $n = 2, \dots, 5$ и двух типов распределений $N(0, 1)$ и $N(0, 0.3)$. Так, для случая, представленного на рис.7, на первый вход сети подаются величины имеющие распределения с параметрами соответственно $\mu = -0.5477, \sigma = 0.7477$ и $\mu = -0.1643, \sigma = 0.2243$, а на второй - с $\mu = 0.5477, \sigma = 0.7477$ и $\mu = 0.1643, \sigma = 0.2243$. Таким образом, подаваемые на каждый из входов распределения имеют сдвиг по среднему величиной $d = 0.3834$.

Таблица 1

n	2	3	4	5
Δ	0.5422	0.8459	1.0980	1.3160

При переходе к большим выборкам n величина d растет, а кроме того, растет и величина расхождения Δ средних значений результирующих n -мерных распределений. В таблице 1 приведены результаты вычислений величины Δ для разных объемов эмпирической выборки $n = 2, \dots, 5$ и двух типов распределений $N(0, 1)$ и $N(0, 0.3)$.

Таким образом, в случае упорядоченной выборки на m -ый вход сети ($m = 1, 2, \dots, n$) подаются хорошо разделяемые распределения⁵. Эта особенность изучаемого преобразования является еще одной причиной того, что сеть значительно быстрее "устанавливает" искомую разделяющую границу.

Следует отметить, что сокращение "площади" разделяющей гиперповерхности, приводя к увеличению скорости обучения и улучшению качества распознавания, имеет также и другое важное следствие, состоящее в том, что для достижения примерно одинакового уровня распознавания для упорядоченной выборки необходимо существенно меньшее число нейронов в скрытом слое. Этот факт становится понятным, если воспользоваться упрощенной моделью сети со ступенчатой переходной функцией, согласно которой каждый нейрон отвечает отдельной гиперплоскости, аппроксимирующей часть разделяющей границы. Выше было показано, что при подаче на вход сети случайной выборки решающая граница представляет собой гиперсферу определенного радиуса. Легко показать, что в случае упорядоченной выборки указанная граница представляет собой часть той же сферы с площадью поверхности в $n!$ раз меньшей, поэтому в предельном случае достаточно большого числа аппроксимирующих гиперплоскостей соотношение числа нейронов скрытого слоя, определяемого отношением площадей решающих границ, также равно $n!$. С другой стороны, оптимальное количество нейронов скрытого слоя для MLP, оцениваемое как $\sim a^{n-1}$ [12], на практике не очень велико, и поэтому для реальной сети преимущество в числе нейронов не столь значительно.

Заключение

В работе рассматриваются некоторые аспекты выбора входных данных для многослойного перцептрона при решении задач классификации событий, представляемых эмпирическими выборками конечного объема в соответствии с их принадлежностью парциальным распределениям, образующих эксперименталь-

⁵ В случае эмпирической выборки на каждый вход подаются распределения с совпадающими средними для обоих классов

ный спектр. Показано, что преобразование входных данных для MLP к вариационному ряду, предложенное в работе [4], приводит к заметному ускорению процесса обучения сети, а также улучшению качества распознавания. Дается теоретическое обоснование изменениям в работе сети при переходе в пространство новых переменных с точки зрения байесовской классификации с минимальным уровнем ошибки и особенностей статистических распределений элементов вариационного ряда. Показано также, что такая форма представления данных, кроме того, позволяет сократить количество нейронов в скрытом слое без потери в точности классификации.

Эта работа поддержана комиссией Европейского Сообщества в рамках сотрудничества ЕС - Россия согласно контракту No. ECRU002.

Литература

- [1] B.Denby: "Tutorial on Neural Networks Applications in High Energy Physics: The 1992 Perspective". In Proc. of the Second International Workshop on "Software Engineering, Artificial Intelligence and Expert Systems in High Energy Physics". January 13-18, 1992 L'Agelaude France-Telecom La Londe-les-Maures (France). New Computing Techniques in Physics Research II, edited by D.Perret-Gallix, World Scientific, 1992, p.287.
- [2] C.Peterson: "Neural Networks in High Energy Physics - Algorithms and Results". Ibidem, p.327.
- [3] E.Rosenblat: "Principles of Neurodynamics". Washington, D.C.: Spartan Books, 1962.
- [4] V.V.Ivanov: "Multidimensional Data Analysis Based on Nonparametric Integral Statistics and Multilayer Perceptrons". JINR Preprint, P10-93-348, Dubna, 1993 (in Russian). Submitted to the "Nucl. Instr. and Meth. in Phys. Res. A.
- [5] П.В.Зрепов, В.В.Иванов, "Интегральные непараметрические статистики $\omega_n^k = n^{\frac{k}{2}} \int_{-\infty}^{\infty} [S_n(x) - P(x)]^k dP(x)$ и их основные свойства. Алгебраический вид, функции распределения и критерии согласия". Сообщение ОИЯИ, P10-92-461, Дубна, 1992.
- [6] D.E.Rumelhart, G.E.Hinton, R.J.Williams: "Learning Internal Representations by Error Propagation" in D.E.Rumelhart & J.L.McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Vol. 1: *Foundations*. MIT Press, 1986.

- [7] L.Lönnblad, C.Peterson and T.Rögnvaldsson, "*Pattern Recognition in High-Energy Physics with Artificial Neural Networks: JETNET 2.0*", *Comp. Phys. Commun.*, **70**, 1992, p.167.
- [8] C.Peterson and T.Rögnvaldsson, L.Lönnblad, "*JETNET 3.0 - A Versatile Artificial Neural Network Package* ", LU TP 93-29, CERN-TH.7135/94, December 1993.
- [9] В.С. Королюк, Н.И. Портенко, А.В. Скороход, А.Ф. Турбин, "*Справочник по теории вероятностей и математической статистике*", "Наука", Москва, 1985.
- [10] С.Е. Clark and G.T. Williams, "*Distributions of the members of an ordered sample*", *Ann. Math. Statist.*, 1958.
- [11] Р. Дуда, П. Харт, "*Распознавание образов и анализ сцен*", "Мир", Москва, 1976.
- [12] T. Rögnvaldsson, "Pattern Discrimination Using Feedforward Networks: A Benchmark Study of Scaling Behavior", *Neural Comput.* **5**, 483 (1993).
- [13] Г.Е. Шиллов, "*Математический анализ. Конечномерные линейные пространства*", "Наука", Москва, 1969.

Рукопись поступила в издательский отдел
9 сентября 1994 года.